



PB94-103496

NTIS[®]
Information is our business.

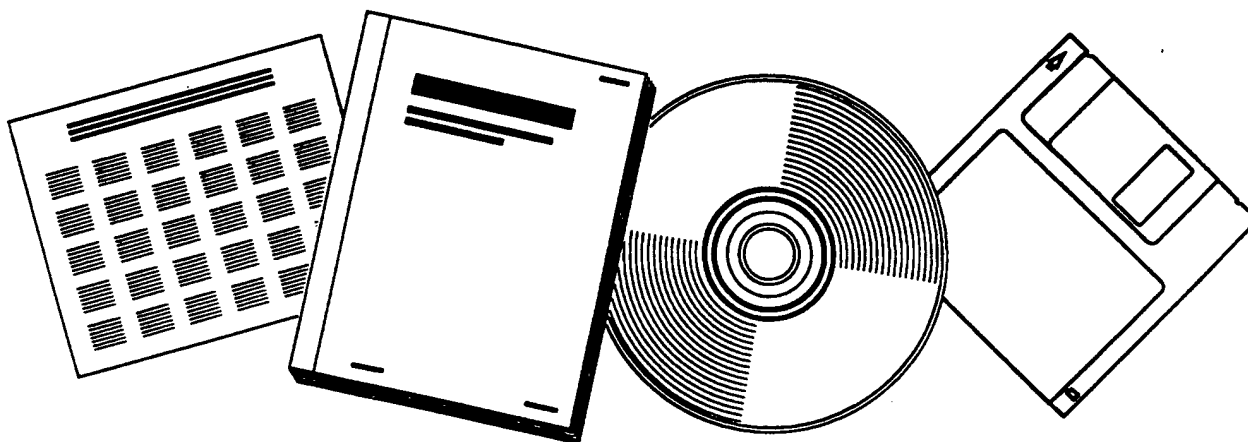
PROTEIN STRUCTURE PREDICTION BY MEMORY-BASE REASONING

19970623 131

THINKING MACHINES CORP
CAMBRIDGE, MA

DTIC QUALITY IMPROVED C

DEC 88



U.S. DEPARTMENT OF COMMERCE
National Technical Information Service

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited



Protein Structure Prediction by Memory-base Reasoning*

Xiru Zhang David Waltz

Thinking Machines Corporation

245 First Street, Cambridge, MA 02142-1214

Brandeis University

415 South Street, Waltham, MA 02254-9110

Jill Mesirov

Thinking Machines Corporation

December 14, 1988

RL88-3,

Abstract

Memory-based reasoning (MBR) is a technique that makes intensive use of memory to recall some specific episodes from the past for problem solving. It is used in this research to predict protein structures based on 112 known structures selected from the Brookhaven Protein Databank. The ϕ and ψ angles of each amino acid in a protein are used to represent its 3-D structure. For this particular problem, we extend MBR to include a recursive procedure to refine its initial prediction and a varying "window" size to take into account the interaction between amino acids apart from different distances along the amino acid sequence. The system implemented, *PHI-PSI*, has been tested with all the available data. It does better than distribution-based guesses for most of the ϕ and ψ angle values.

1 Motivation and Introduction

When faced with a problem, what should we do if we do not have enough domain knowledge ("rules") to solve it but do have a set of examples? Problems of this kind abound in our daily life as well as in scientific investigations. In this research we take the protein structure prediction problem as our vehicle of investigation, and focus on the development of a computational model for solving such problems.

*This work was supported in part by the Defense Advanced Research Projects Agency, administered by the U.S. Air Force Office of Scientific Research under contract number F49620-88-C-0058.

It is known that all proteins in all species, from bacteria to humans, are composed of the same set of 20 amino acids and that every protein has a unique amino acid sequence which specifies its three-dimensional structure. It is now fairly easy to determine a protein's amino acid sequence or even to chemically synthesize a new sequence, but extremely difficult to determine its 3-D structure.¹ Automatic determination of protein structure is of great scientific and practical value because it is closely related to protein function. Though we know the amino acid sequences for thousands of proteins, we only know the structures for a few hundred of them. And the principles underlying the correspondence between the structure and the amino acid sequence are poorly understood. An interesting question is: How do we use the information hidden in the known structures to help predict the unknown?

Memory-based reasoning (MBR) [14] assumes that *the intensive use of memory to recall specific episodes from the past should be the foundation of machine reasoning*. It can make use of massively parallel hardware such as the Connection Machine [5] to produce an efficient implementation. Given a problem, an MBR system "recalls" all the precedents in its memory that bear some resemblance to the problem and derives a solution based on those retrieved precedents through some decision-making process. Based on this idea, we developed the system *PHI-PSI* which makes use of known protein structures to predict the structure of proteins of which we only know the amino acid sequences.

In this paper we discuss the design principles and initial results of *PHI-PSI*. We also include brief background information on the biology needed to understand this application (some of which is put in footnotes). We conclude with a discussion about the related work and plans for future research.

2 Amino Acids and Their ϕ and ψ Angles

Proteins are composed of linear sequences of amino acids. There are 20 different types of amino acids, they are the "alphabet" of all the proteins. The 20 amino acids differ from each other only in their side-chains (a group of atoms) which determine their physical properties (see figure 1 (left)). According to these properties, amino acids can be classified into similar/dissimilar groups, such as small/large, polar/non-polar, hydrophobic/hydrophilic, etc. In a protein, the carboxyl group of one amino acid is joined to the amino group of another amino acid by a *peptide bond*. Many amino acids, usually a hundred or more, are joined by peptide bonds to form a *polypeptide chain* (also called a *primary sequence*; each amino acid in it is sometimes called a *residue*), as shown in figure 1 (right). Figure 2 (left) shows the spatial relations for atoms joined by two peptide bonds. The atoms between two C_α 's are basically on a plane and the bond lengths are essentially fixed.

¹ Right now mainly through crystallography, which is at best time consuming - typically ten man-years per structure, and at worst impossible because crystals are either unsuitable or unavailable.

There are two degrees of freedom about the relative positions of two adjacent planes – the ϕ and ψ angles – which are the primitive structural descriptors used in this work.²

3 Approach

PHI-PSI was developed based on the hypothesis that if two amino acids have similar physical properties and they occur in a similar physical environment, then they should form similar structures (e.g. similar ϕ - ψ angles). Here are some terms used in the following discussion: (1) **test protein** – the protein whose structure is going to be predicted, and whose amino acid sequence is the input to *PHI-PSI*; (2) **database** – all the protein data we have except the test protein; here both the amino acid sequences and the ϕ and ψ angles of each amino acid are known; (3) **window** – one-dimensional frame of slots that is to be overlaid on and moved over a test protein sequence to access segments of its amino acids.

3.1 The Basic Algorithm

For each test protein, *PHI-PSI* works as follows:

Step 1. Specify the initial parameters, such as the initial window size W , the window weight pattern P (there is a weight associated with each slot in the window), and N , the number of best matches to keep (from which a prediction is made), etc. (These parameters will be discussed in more detail below.)

Step 2. Move the window over the test protein, and at each position, extract an amino acid segment S of length W , and do:

1. move the same window over all the protein sequences in the database and generate all the possible amino acid segments s_i of length W , $i = 1, 2, \dots, m$;
2. match S against all s_i , $i = 1, 2, \dots, m$, and compute a score using a scoring function which will be described in the next section;
3. select the N segments from $\{s_1, \dots, s_m\}$ which have the highest N scores. The prediction of the ϕ and ψ angles of S 's centermost amino acid is made by majority of the ϕ and ψ values of the amino acids in the N selected segments.

Step 3. If the recursive mode is chosen, adjust the parameters (e.g. the window size) and repeat Step 2 unless the end conditions are met or *PHI-PSI* has gone through a pre-specified number of recursive levels.

² Most work in protein structure prediction has focused on *secondary structures*, which refer to the regular, repetitive spatial arrangements of residues that are close to one another in the polypeptide chain. They include: (i) α *helix*, where the polypeptide chain has a helical shape to produce a rodlike structure; (ii) β *sheet*, where the polypeptide chain is almost fully extended; (iii) β *turn*, where within a few (usually 4) residues the polypeptide chain turns $\sim 180^\circ$. *Coil* is often used to mean "none of the above." We use the ϕ - ψ angles instead in our research for several reasons: (1) there are no agreed assignments of the secondary structures in certain cases; (2) ϕ - ψ angles compose a richer vocabulary; for example, some biologists have found 8 types of β turns, which can not be distinguished if one just calls them all turns; (3) ϕ - ψ angles can be used to describe any part of a protein, not just the secondary structures.

3.2 The Similarity Measure

Given two patterns, a scoring function computes a score to represent how similar they are to each other. To define a scoring function, several factors need to be considered: (a) the similarity matrices, which specify how similar the primitive components (here, the amino acids) are to each other in terms of some properties (usually one matrix for each property); (b) the weight of each similarity matrix, which represents how important the corresponding property is to the overall similarity; (c) the weights associated with each position of the pattern, which indicates how important that position is to the whole pattern; and (d) if the matching is done recursively, how strongly the previous result should affect the next match. The following is the function used by *PHI-PSI*, which takes two segments of amino acids $X = X_1 X_2 \dots X_n$, $Y = Y_1 Y_2 \dots Y_n$ as arguments:

$$Score = \sum_{i=1}^n Wp_i \cdot \left\{ \sum_{j=1}^m Wm_j \cdot S_j[X_i, Y_i] - Wa_i \cdot \max(|X_i[\phi] - Y_i[\phi]|, |X_i[\psi] - Y_i[\psi]|) \right\}$$

where X_i and Y_i are the amino acids in X and Y respectively, $i \in [1, n]$; $X_i[\phi]$ means the ϕ angle of X_i , whose value is within $[-180, 180]$; Wp_i is the window slot weight for position i ; S_j is the j th similarity matrix, $S_j[X_i, Y_i]$ means the entry value for amino acid pair (X_i, Y_i) ; Wm_j is the weight for S_j ; Wa_i is the weight for previously-predicted ϕ and ψ angle values, which is zero when the angles are unknown.

What this function does is the following: for every pair (X_i, Y_i) from X and Y , it computes the weighted sum of the corresponding entries for X_i and Y_i in all the similarity matrices, and if in recursive mode, subtracts the difference between the previously-predicted and currently-retrieved ϕ - ψ values. This sum is taken as the *subscore* for each pair (X_i, Y_i) . The function then computes the weighted sum of all the subscores as the score of matching the two segments X and Y . So, the matrices for important properties should have more weight, and the important positions in the window should have more weight in order for the scores to reflect the structural similarity between the two amino acid segments.

We have used 10 amino acid properties, such as *size*, *hydrophobicity*, *polarity*, etc.[16] A 20×20 similarity matrix is computed for each property based on values obtained from biology literature. Each matrix entry has a value between $[0, 1]$, representing the degree of similarity between a pair of amino acids in terms of that property. The diagonal elements of each matrix, that is, the entry for pair (A_i, A_i) for some amino acid A_i , should have value 1.0. However, since we do not fully understand the similarity between amino acids in terms of forming protein structures, exact matches are always preferred. To emphasize this, the diagonal elements are increased to 1.5.⁴

³ Note that -180° is equal to $+180^\circ$ here.

⁴ We tested various values between $[1, 2]$, 1.5 seems to give the best prediction.

3.3 The Window Size

There is a trade-off between the amount of information and the level of noise when we choose a window size. The larger the window, the more information it contains, but when the window is too large, the matching process may be misled by the "noise," i.e., the irrelevant sequentially distant information.

The technique used in this work to make the trade-off is to start with a small window, and to increase the window size gradually. At each level of recursion, the previous prediction from a smaller window is used in finding the best matches for the next prediction of a larger window. This way, we can catch the information of the very short range interactions as well as take into account the somewhat longer range interactions between the amino acids in a primary sequence.

To a first approximation, we can assume that the structure of a protein is solely determined by the interactions among its amino acids, and the amino acids interact with each other only if they are close to each other in space. For every position i in a primary sequence, we computed the probability (approximated by the frequency) by which the residues at position $i \pm 1$, $i \pm 2$, ... are within 7\AA from residue i .⁵ Residues two positions or less apart on the amino acid sequences are almost always close to each other, this is in accordance with the standard bond length and configuration; residues within four positions are close to each other much more often than are the rest. Thus either 5 or 7 can be used as the initial window size.

3.4 The Weight Pattern

The weight pattern for the window represents the influence that the amino acids at different positions have on the structure of the amino acid in the center of the window. So if position i is at the center of the window, the weight of position $i \pm j$ should be in proportion to the probability by which residue $i \pm j$ stay close to residue i (based on the assumption in the last section). We have tested different window weight patterns; a typical one looks like:
... 3 3 4 4 5 4 4 3 3

3.5 Find the "Allies" for the Best Matches

After we find the top N matches (the ones with the highest N scores) from the database for an input amino acid segment, we need a way to make a decision about which ϕ and ψ values to use as predictions for the input.

First, why do we bother to use the top N matches, not just the one with the highest score? Two reasons: (1) Usually the top matches have very similar scores. Recall that the scoring function is based on quite a few factors. We are not quite sure about the exact

⁵ 7\AA is roughly the distance within which two residues can directly interact with each another.

weights for these factors. The actual values used are at best approximations of the optimal values. So a small difference in the score may not mean much. (2) Even if two (short) segments of amino acids match exactly, they do not necessarily form the same structure in two proteins. However, if among the top N matches, the majority of them have a similar structure, then the input will at least have the *tendency* to form that structure also.

PHI-PSI makes this decision in the following way: for each segment in the top N matches it finds out how many other segments in this group have similar values for the centermost residue; these are called its "allies" (the threshold for "similar values" here is another parameter that can be adjusted). The one with the largest number of allies is chosen as the prediction.

4 Discussions

4.1 Selection of Data

Homologous proteins⁶ have similar amino acid sequences and structures. So if two proteins in the database are homologous, then when we try to predict the structure of one of them, we almost always find the best match from the other. This way, though we can have a high prediction accuracy, the result is deceiving.⁷ We used a sequence comparison package developed on the Connection Machine [8] to find the homologous protein clusters and removed all but one sequence (usually the longest one and/or the one with the highest resolution) from each cluster. The 112 amino acid sequences left have been used in this work.

4.2 Initial Results and Analysis

We made one complete run of *PHI-PSI* on the whole database, that is, for every amino acid sequence p in the database:

- select p as test protein and use the rest as known proteins;
- run *PHI-PSI* to predict p 's structure;
- compare the prediction with p 's real structure and compute the prediction errors.

The parameters used: *initial_window_size* = 5; *recursive_level* = 5.

This involves a large amount of computation. For example, there are 112 primary sequences in our database, which contain 18713 amino acids altogether; the recursive level is set to 5 and at each level the window size is increased by 2, so on average there are 10 amino acids in each segment; each amino acid sequence generates $(18713 \div 112) - 10 \approx 157$ segments; there are 10 property matrices; for each amino acid segment in a test protein,

⁶ Proteins that have common ancestors.

⁷ Given a protein with unknown structures, it would be very helpful if we know the structure of a protein that is homologous to it. But in most cases, we do not know any of its homologous proteins.

PHI-PSI matches it against the whole database (except itself); thus the total number of table lookups (i.e. calls to the similarity matrix entries) for a complete run is:

$$(112 \times 157) \times [(112 - 1) \times 157 \times 10 \times 10 \times 5] = 153218184000 \approx 1.5 \times 10^{11}.$$

There need to be roughly the same amount of multiplications (see the scoring function in section 3.2). The complete run took about 100 hours on a 4K Connection Machine without floating point processors (FPPs). With FPPs and the new indirect-addressing facility for virtual processors, this computation should be 4 times as fast.⁸ Also, the algorithm speeds up linearly with the number of processors.

The prediction errors are computed in terms of ϕ and ψ angles. There are several ways to measure the errors, such as: (1) **residue errors** – the difference between the real angle values computed from the 3-D coordinates and the values predicted by the algorithm for a particular residue in a protein; (2) **overall errors** – the average of the residue errors of all the proteins in the database.

4.2.1 The Overall Errors

The overall errors of the complete run are: $\phi_error = 37.7^\circ$, $\psi_error = 63.7^\circ$. For comparison, the average differences of ϕ and ψ angle values among 1000 randomly selected residues in the database are: *random- ϕ -difference* $\approx 56^\circ$, *random- ψ -difference* $\approx 89^\circ$. The prediction errors are considerably smaller than the random differences,⁹ which demonstrates that our algorithm can really find some correspondences between segments of amino acids and their structure.¹⁰

4.2.2 The Residue Errors

Figure 3 shows the average residue errors (the dark curves) and the “random differences” (the thin lines) for all the angle values. The random difference for angle value V ($V \in [-180, 180)$) is computed by randomly picking up 1000 residues from the database and calculating the average difference between V and these angle values. A comparison of the prediction error curve with the random difference curve can give us an idea about how well *PHI-PSI* does with each particular angle value. Put another way, for each residue in a test protein, the “error” of a random guess (based only on the distribution of the ϕ - ψ angle values in the database) has the greatest probability to fall on the random difference curve. So the vertical distance between the random difference curve and the prediction error curve shows how much better *PHI-PSI* does than distribution-based guessing.

⁸ This estimate is based on most multiplications being carried out on 20-bit integers.

⁹ The “random differences” are not totally random, they are determined by the distribution of ϕ - ψ values in the database.

¹⁰ Notice that with *resolution* = 2.5Å, the ϕ and ψ angles can only be measured with accuracy around $20^\circ \sim 30^\circ$. Quite a few protein structures in our database have that resolution.

From figure 3 we can see that *PHI-PSI* could do pretty well for ϕ around -60° and for ψ around -40° and 110° . *PHI-PSI* did not do very well for ϕ in region $[130, 170]$ and for ψ in regions $[-170, -140]$ and $[70, 90]$. In figure 2 (right) we plotted the distribution of all the ϕ and ψ angle values in our database (which is called *Ramachandran plot*). We can observe the following phenomena: *PHI-PSI* did much better in dense ϕ - ψ regions than in sparse ϕ - ψ regions. Recall that our prediction algorithm works by finding similar segments of amino acids for each one in a test protein. When the ϕ - ψ angles of an amino acid in the test protein is in a dense region, there are a lot of precedents in the database, thus *PHI-PSI* is likely to find ones similar to it; when it is in a sparse region, there are simply not enough precedents for a good prediction. The ϕ and ψ angles of α helix and β sheet¹¹ also fall into the dense regions, so another explanation for the observed phenomena is that there is a closer correlation between amino acid segments and their structure for helices and sheets than for other parts of a protein.

5 Related Work

The idea of memory-based reasoning is related to the theories on *analogy* [17], *dynamic memory* [13], *case-based reasoning* [7] [3], and *instance-base reasoning and learning* [6]. One common theme among all these theories is that one can solve a problem by recalling one or more related precedents and deriving a solution based on them.

Most heuristic methods for protein structure prediction have focused on the secondary structures and have adopted a "local approach," i.e., to predict a residue's structure based on its neighboring residues along the primary sequence. Chou & Fasman [1] used the frequencies by which each amino acid appears in α helix and β sheet to predict the helices and sheets in a new protein. Garnier *et al.* [2] computed the correlation of residue $j + m$ and the "state" of residue j (one of $\{\text{helix}, \text{sheet}, \text{coil}\}$), for $m \in [1, 8]$, and used this information in prediction. Rooman & Wodak [12] developed a set of "sequence motifs" (amino acid patterns) for predicting the secondary structures, and their conclusion is that the identification of predictive sequence motifs is limited by the size of the currently available data. There have been several attempts to use the known data directly in prediction. Levin *et al.* [10] used a window of length 7 to predict the secondary structures of every amino acid inside the window. Sweet [15] used a window of length 12 and made use of the probability distribution of ϕ - ψ angles to predict the secondary structures. These direct methods, including our method, have the advantage that their prediction accuracies are likely to increase as the size of the available data increases. Qian & Sejnowski [11] applied the back-propagation algorithm to the prediction of α helix and β sheet. They got $\sim 64\%$ accuracy for 15 test sequences, which they claimed is the best result so far.

A few AI researchers have attached the 3-D structure prediction problem. Hayes-

¹¹ See footnote 2 for definitions.

Roth *et al.* [4] used multiple sources of information (mainly through nuclear magnetic resonance (NMR); they also assumed that a protein's secondary structures are known) and the blackboard control architecture to identify legal positions for each of a protein's constituent structures (atoms, amino acids, helices, *etc.*), which usually results in a large number of possible structures. Currently the NMR techniques are limited to proteins with < 100 residues. The ARIADNE system [9] is another interesting effort. It is essentially a recognition system that uses hierarchical representation of protein structures to decide whether a primary sequence can fold into a given 3-D structure.

6 Summary and Future Research

This paper reports our initial work on the prediction of protein structures by Memory-based Reasoning. There has not been much work done in predicting the ϕ - ψ angles directly from the known data. We feel that this is an important direction to explore, and that Memory-based Reasoning is the right technique for this task. We will continue our research in several directions, among them are:

- (1) Currently in *PHI-PSI* when we compute the weighted sum of matrix entries and the weighted sum of subscores, all the segments in the database share the same two sets of weights: one set associated with the similarity matrices, the other associated with the window. This may not be the optimal way to do it, because the "important properties" and the "important positions" may be different in each case, and even the window size for each case does not necessarily have to be the same. We are developing learning algorithms to automatically identify these optimal parameters for each case.
- (2) We have used a similar network structure as in [11] to predict the ϕ - ψ angles of a few sequences, and found the overall accuracy was very close to that of *PHI-PSI*'s. However, their predictions for each particular sequence are not the same, which suggests that the combination of multiple methods may give better results than any single one. We plan to implement a system which uses multiple methods and to derive a way to combine their results by carefully examining under what conditions each method is likely to give correct prediction.
- (3) Amino acids in a protein interact with others that are close to them in space, some of which may be far away along the primary sequence. It has been known for a long time that this kind of "global interaction" is crucial for protein folding and structural stability. One way to take into account this global interaction is to use the information about the super-secondary structures.¹² For future research, in *PHI-PSI*, each known amino acid sequence will be associated with not only the ϕ and ψ angles, but also the secondary and

¹² A super-secondary structure is a group of secondary structures that are arranged in some more or less regular way.

super-secondary structure information. *PHI-PSI* will include this higher-order structure information in finding best matches recursively.

7 Acknowledgment

Many thanks to Robert Jones, who read the draft very carefully and made many suggestions; to Gerald Fasman, who supplied us with a lot of information and the data. The comparison with back-propagation algorithm was done jointly with Dean Pomerleau. Discussions with Eric Lander were also helpful.

References

- [1] Peter Y. Chou and Gerald D. Fasman. Prediction of Protein Conformation. *Biochemistry*, 13(2), 1974.
- [2] J. Garnier, D. J. Osguthorpe, and B. Robson. Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins. *Journal of Molecular Biology*, 120, 1978.
- [3] K. J. Hammond. CHEF: A model of Case-based Planning. In *AAAI-86*, pages 267 - 271, 1986.
- [4] Barbara Hayes-Roth et al. PROTEAN: Deriving Protein Structure from Constraints. In *AAAI-86*, 1986.
- [5] W. Daniel Hillis. *The Connection Machine*. The MIT Press, 1985.
- [6] Dennis Kibler and David W. Aha. *Instance-Based Prediction of Real-Valued Attributes*. Technical Report 88-07, University of California, Irvine, March 1988.
- [7] J. Kolodner, R. Simpson, and K. Syrcara-Cyranski. A Process Model of Case-Based Reasoning in Problem Solving. In *Proceedings of the Ninth IJCAI*, pages 284 - 290, 1985.
- [8] Eric Lander, Jill P. Mesirov, and Washington Taylor IV. Protein Sequence Comparison on a Data Parallel Computer. In *Proceedings of the 1988 International Conference on Parallel Processing*, pages 257 - 263, 1988.
- [9] Richard H. Lathrop, Teresa A. Webster, and Temple F. Smith. Ariadne: Pattern-directed Inference and Hierarchical Abstraction in Protein Structure Recognition. *CACM*, 30(11), 1987.
- [10] J. M. Levin, B. Robson, and J. Garnier. An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS*, 205(2):303 - 308, 1986.
- [11] Ning Qian and Terrence J. Sejnowski. Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *Journal of Molecular Biology*, 202, 1988.
- [12] Mariann J. Rومان and Shoshana J. Wodak. Identification of predictive sequence motifs limited by protein structure data base size. *Nature*, 335(1):45 - 49, September 1988.
- [13] Roger C. Schank. *Dynamic Memory*. Cambridge University Press, 1982.
- [14] Craig Stanfill and David Waltz. Toward Memory-based Reasoning. *CACM*, 29(12), 1986.

- [15] Robert M. Sweet. Evolutionary Similarity Among Peptide Segments Is a Basis for Prediction of Protein Folding. *Biopolymers*, 25:1565 – 1577, 1986.
- [16] William R. Taylor. The Classification of Amino Acid Conservation. *Journal Theoretical Biology*, 119:205 – 218, 1986.
- [17] Patrick H. Winston. Learning and Reasoning by Analogy. *CACM*, 23(12):689 – 703, 1980.

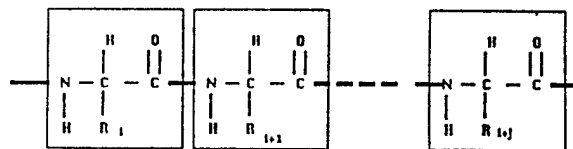
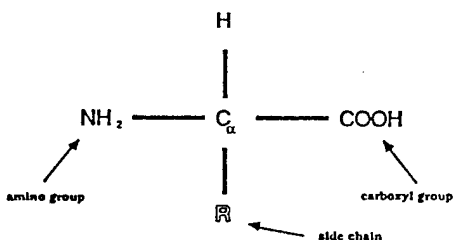


Figure 1: *Left* – The chemical structure of an amino acid. *Right* – An amino acid sequence.

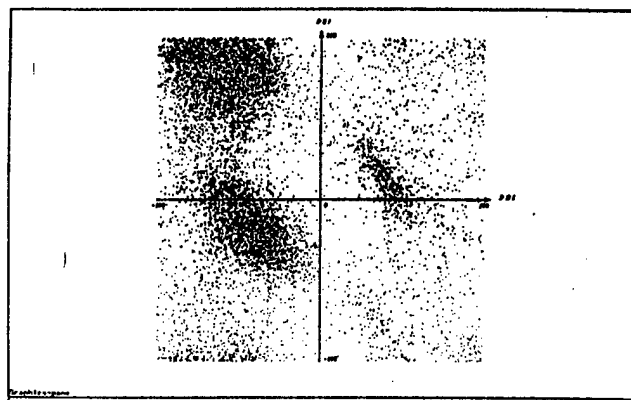
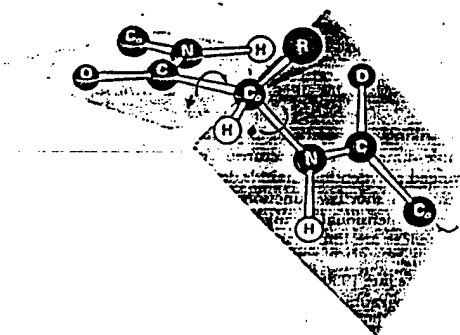


Figure 2: *Left* – The ϕ and ψ angles. *Right* – The distribution of ϕ and ψ angle values in our database.

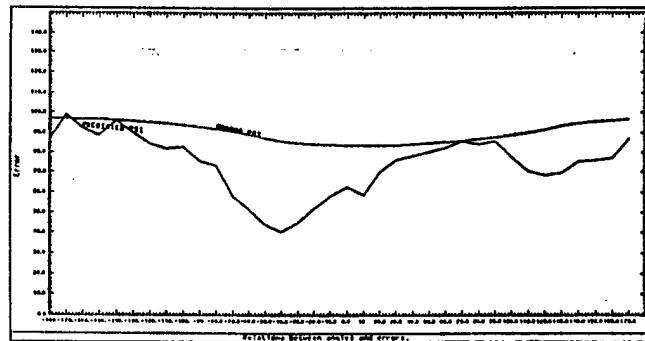
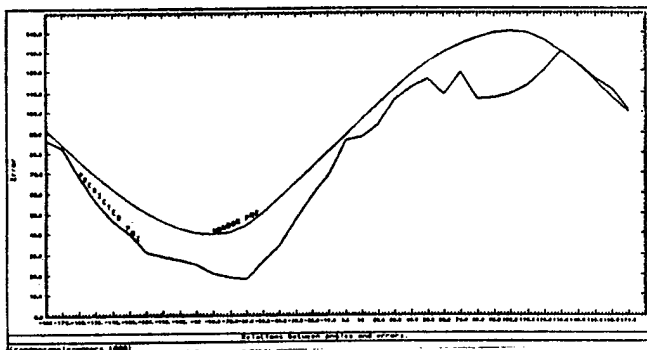



Figure 3: *Left* – Residue error curve for ϕ . *Right* – Residue error curve for ψ .

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
<small>Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.</small>				
1.  PB94-103496		2. REPORT DATE 14, December 88		
		3. REPORT TYPE AND DATES COVERED Technical		
4. TITLE AND SUBTITLE Protein structure prediction by memory-base reasoning			5. FUNDING NUMBERS DARPA F49620-88-C-0058	
6. AUTHOR(S) Zhang, X. Waltz, D and Mesirov, J.				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Thinking Machines Corp. 245 First Street Cambridge, MA 02142-1264			8. PERFORMING ORGANIZATION REPORT NUMBER TMC-155	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) 1400 Wilson Boulevard, Arlington, VA 22209-2308			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION / AVAILABILITY STATEMENT			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) Memory-based reasoning is a technique that makes intensive use of memory to recall some specific episodes from the past for problem solving. It is used in this research to predict protein structures based on 112 known structures selected from the Brookhaven Protein Data bank. The angles of each amino acid in a protein are used to represent its 3-D structure. We extend MBR to include a recursive procedure to refine its initial prediction and a varying "window" size to take into account the interaction between amino acids apart from different distances along the amino acid sequence. The system implemented, PHI-PSI, has been tested with all the available data. It does better than distribution-based guesses for most of the angle values.				
14. SUBJECT TERMS Reasoning and Learning			15. NUMBER OF PAGES 13	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE SAR	19. SECURITY CLASSIFICATION OF ABSTRACT SAR	20. LIMITATION OF ABSTRACT SAR	

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to stay *within the lines* to meet optical scanning requirements.

Block 1. Agency Use Only (Leave blank).

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C - Contract	PR - Project
G - Grant	TA - Task
PE - Program Element	WU - Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. (If known)

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."

DOE - See authorities.

NASA - See Handbook NHB 2200.2.

NTIS - Leave blank.

Block 12b. Distribution Code.

DOD - Leave blank.

DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.

NASA - Leave blank.

NTIS - Leave blank.

Block 13. Abstract. Include a brief (*Maximum 200 words*) factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code (*NTIS only*).

Blocks 17 - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

NTIS does not permit return of items for credit or refund. A replacement will be provided if an error is made in filling your order, if the item was received in damaged condition, or if the item is defective.

Reproduced by NTIS

National Technical Information Service
Springfield, VA 22161

*This report was printed specifically for your order
from nearly 3 million titles available in our collection.*

For economy and efficiency, NTIS does not maintain stock of its vast collection of technical reports. Rather, most documents are printed for each order. Documents that are not in electronic format are reproduced from master archival copies and are the best possible reproductions available. If you have any questions concerning this document or any order you have placed with NTIS, please call our Customer Service Department at (703) 487-4660.

About NTIS

NTIS collects scientific, technical, engineering, and business related information — then organizes, maintains, and disseminates that information in a variety of formats — from microfiche to online services. The NTIS collection of nearly 3 million titles includes reports describing research conducted or sponsored by federal agencies and their contractors; statistical and business information; U.S. military publications; audiovisual products; computer software and electronic databases developed by federal agencies; training tools; and technical reports prepared by research organizations worldwide. Approximately 100,000 *new* titles are added and indexed into the NTIS collection annually.

For more information about NTIS products and services, call NTIS at (703) 487-4650 and request the free *NTIS Catalog of Products and Services*, PR-827LPG, or visit the NTIS Web site
<http://www.ntis.gov>.

NTIS

***Your indispensable resource for government-sponsored
information—U.S. and worldwide***

Thank you for your order from NTIS!

The products and or services listed here may also be of interest to you.



SCIENCE

**Science for all Children: A Guide to Improving
Elementary Science Education in Your School District
and
Resources for Teaching Elementary School Science**
National Science Resource Center

Today's children start school with plenty of knowledge—TV, the Internet and other virtual reality systems have opened their minds to a whole world of opportunities. But there is a difference between watching the launch and riding in the spaceship—manning the remote is not the same as manning the instrument panel.

Through a new approach of inquiry-based, hands-on science, even kindergarten children can explore how the world works. They learn to ask questions, conduct experiments, use tools, interpret data and communicate ideas. Studies show children develop independent thinking and problem-solving skills in an activity-related environment, rather than the traditional passive learning environment of textbooks and memorizing facts. *Science for All Children* explains the rationale for inquiry-based science and provides guidelines for planning such a program at the elementary school level. It presents a collection of case studies that show how models are being implemented in school districts nationwide—i.e., alliances with scientists and engineers from corporations and academic institutions. Teachers, administrators, scientists, and parents who want to expand the sphere of science education in their schools will want this document.

NTIS is distributing these new guides from the National Science Resource Center, which is a partnership between the Smithsonian Institution and the National Academy of Sciences dedicated to improving science education in schools. These documents are available at a special price when you order both.

To order the set, use *Order Number PB97-163109LPH*

\$79.50 plus handling fee. Orders outside the U.S., Canada, and Mexico \$159 plus handling fee.

Science for all Children: A Guide to Improving Elementary Science Education in Your School District

Order Number: PB97-138010LPH

\$44 plus handling fee. Orders outside the U.S., Canada, and Mexico \$88 plus handling fee.

Resources for Teaching Elementary School Science

An excellent reference guide to 350 inquiry-based curriculum packages.

Order Number: PB96-184254LPH

\$49 plus handling fee. Orders outside the U.S., Canada, and Mexico \$98 plus handling fee.

ENERGY

**U.S. Nuclear Regulatory Commission Regulatory Guide 1.160, Revision 2,
Monitoring the Effectiveness of Maintenance at Nuclear Power Plants**

U.S. Nuclear Regulatory Commission, Office of Nuclear Regulatory Research, Washington DC

The *U.S. Nuclear Regulatory Commission Regulatory Guide* series makes available to the public methods of implementing specific parts of the commission's regulations. It describes techniques used by the NCR in evaluating specific problems, and it provides guidance to applicants who are involved with nuclear reactors.

This guide is available as an ongoing subscription. Call the NTIS Subscriptions Department at (703) 487-4630 for pricing.

Order number: PB97-926501LPH (for single issue)

\$10 plus handling fee. Outside the U.S., Canada, and Mexico \$20 plus handling fee.

Prices are subject to change.

NTIS Sales Desk: (703) 487-4650

TRANSPORTATION

Navigational and Vessel Inspection Circulars and the Merchant Vessels of the United States (on CD-ROM)

U.S. Coast Guard

This new CD-ROM contains a fully searchable set of all 4,000 pages of *Navigational and Vessel Inspection Circulars* published by the U.S. Coast Guard between July 1952 and May 1996 and is available from NTIS. Also included on the CD-ROM is the current USCG database of merchant vessels registered in the U.S.

NVICs are published by the Coast Guard to assist marine safety personnel and the marine industry by clarifying and expanding upon commercial vessel safety requirements. The documents are provided as both text and image files to allow full-text searching as well as viewing, printing, and faxing on any PC running Windows with a CD-ROM reader.

Individual *Navigational and Vessel Inspection Circulars* are available in paper copy from NTIS. For further information, call NTIS at (703) 487-4650 or visit NTIS Web site
<http://www.ntis.gov/business/nvic.htm>.

Order Number: PB97-500664LPH

\$40 plus handling fee.

Orders outside the U.S., Canada, and Mexico \$80 plus handling fee.

ENVIRONMENT

Water Test Methods and Guidance from EPA (on CD-ROM)

U.S. Environmental Protection Agency

EPA's Office of Water has taken the initiative to provide its methods and guidance documents on CD-ROM. The CD-ROM contains more than 330 drinking water and wastewater methods and guidance from over 50 EPA documents including: MCAWW; Metals, Inorganic and Organic Substances in Environmental Samples; 40 CFR Part 136 Appendix A, B, C & D; 500, 600, and 1600 series; Whole Effluent Toxicity Methods.

The CD-ROM contains search and retrieval software and requires WINDOWS 3.1 or greater or Mac 68020 processor or greater.

Order Number: PB97-501308LPH

\$60 plus handling fee.

Outside the U.S., Canada, and Mexico \$90 plus handling fee.

U.S. Industry and Trade Outlook, '98

Table of Contents

1	Metals and Industrial Minerals Mining	26	Information Services
2	Coal Mining	27	Computer Hardware and Equipment
3	Crude Petroleum and Natural Gas	28	Computer Software and Networking
4	Petroleum Refining	29	Space Commerce
5	Electricity Production and Sales	30	Telecommunication Services
6	Construction	31	Telecom and Navigation Equipment
7	Wood Products	32	Entertainment
8	Construction Materials	33	Apparel
9	Electric Lighting and Wiring	34	Footwear, Leather, and Leather Products
10	Textiles	35	Processed Foods and Dairy
11	Paper and Allied Products	36	Motor Vehicles
12	Chemicals and Allied Products	37	Auto Parts and Accessories
13	Plastic and Rubber	38	Household Consumer Durables
14	Metals	39	Recreational Equipment
15	General Components	40	Other Consumer Durables
16	Microelectronics	41	Wholesaling
17	Metalworking Equipment	42	Retailing
18	Production Machinery	43	Transportation
19	Electrical Equipment	44	Travel Services
20	Environmental Technologies	45	Health and Medical Services
21	Aerospace	46	Medical and Dental Instruments and Supplies
22	Shipbuilding and Repair	47	Financial Services
23	Industrial and Analytic Equipment	48	Security and Commodity Futures Trading
24	Photographic Equipment	49	Business Professional Services
25	Print and Electronic Media	50	Education and Training

Place your order today!⇒

U.S. Industry and Trade Outlook '98

Successor to the U.S. Industrial Outlook — the most widely read and respected single source guide to U.S. industry

Content includes:

- 50 chapters covering most important manufacturing and nonmanufacturing sectors!
- New industries not previously covered such as electricity production!
- Expanded coverage in both manufacturing and nonmanufacturing industries!
- Charts for each chapter—provide a quick look at economic and trade trends!

Reserve your copy today!

Be one of the first to receive a copy of this vital reference tool!



Availability: September 1997 Order Number: PB97-165443LPH Price: \$69.95 plus handling fee for U.S., Canada, Mexico

SHIP TO ADDRESS (please print or type)

CUSTOMER MASTER NUMBER (IF KNOWN)		DATE
ATTENTION / NAME		
ORGANIZATION	DIVISION / ROOM NUMBER	
STREET ADDRESS		
CITY	STATE	ZIP CODE
PROVINCE / TERRITORY	INTERNATIONAL POSTAL CODE	
COUNTRY		
PHONE NUMBER ()	FAX NUMBER ()	
CONTACT NAME	INTERNET E-MAIL ADDRESS	

ORDER BY PHONE

(eliminate mail time)
8:30 a.m.—5:00 p.m., Eastern Time, M–F.
Sales Desk: (703) 487-4650
TDD: (703) 487-4639

ORDER BY FAX

24 hours/7 days a week: (703) 321-8547
To verify receipt of fax: (703) 487-4679
7:00 a.m. — 5:00 p.m., Eastern Time, M–F.

ORDER BY MAIL

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161

RUSH SERVICE

(available for an additional fee)
1-800-553-NTIS

NTIS ORDERNOW™ ONLINE

Order the most recent additions to the NTIS collection at NTIS Web site
<http://www.ntis.gov/ordernow>.

ORDER VIA E-MAIL

Order via E-mail 24 hours a day:
orders@ntis.fedworld.gov
If concerned about Internet security, you may register your credit card at NTIS. Simply call (703) 487-4682.

METHOD OF PAYMENT (please print or type)

<input type="checkbox"/> VISA	<input type="checkbox"/> MasterCard	<input type="checkbox"/> American Express
CREDIT CARD NUMBER	EXPIRATION DATE	
CARDHOLDER'S NAME		
<input type="checkbox"/> NTIS Deposit Account Number:		
<input type="checkbox"/> Check/Money Order enclosed for \$ <small> payable to NTIS in U.S. dollars</small>		

PRODUCT SELECTION (please print or type)

NTIS PRODUCT NUMBER	INTERNAL CUSTOMER ROUTING (OPTIONAL) up to 8 characters	UNIT PRICE	QUANTITY	TOTAL PRICE
U.S. Industry and Trade Outlook '98 PB97-165443LPH		\$69.95		\$
TOTAL				\$
HANDLING FEE PER TOTAL ORDER				\$ 4.00
GRAND TOTAL				\$



U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Technical Information Service
Springfield, VA 22161 (703) 487-4650